CUMENTATION PAGE

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1993	Professional Paper

4. TITLE AND SUBTITLE

SAMPLE SIZE DETERMINATION FOR GROUPED EXPONENTIAL OBSERVATIONS: A COST FUNCTION APPROACH

5. FUNDING NUMBERS

PR: ME06
PE: OMN
WU: DN301015

6. AUTHOR(S)

K. J. Lui, D. Steffey, J. K. Pugh

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division
San Diego, CA 92152-5001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Public Health Service
Operational Medicine Informatics
915 Second Avenue
Seattle, WA 98174

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Calculating the required sample size for a desired power at a given type I error level, we often assume that we know the exact time of all subject responses whenever they occur during our study period. It is very common, however, in practice that we only monitor subjects periodically and, therefore, we know only whether responses occur or not during an interval. This paper includes a quantitative discussion of the effect resulting from data grouping or interval censoring on the required sample size when we have two treatment groups. Furthermore, with the goal of exploring the optimum in the number of subjects, the number of examinations per subject for test responses, and the total length of a study time period, this paper also provides a general guideline about how to determine these to minimize the total cost of a study for a desired power at a given α-level. A specified linear cost function that incorporates the costs of obtaining subjects, periodic examinations for test responses of subjects, and the total length of a study period, is assumed, primarily for illustrative purpose.

Published in *Biometrical Journal*, 35(1993)6, pp. 677-688.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Survival Analysis	Logistic Regression	
Trend Test	Forecast Modeling	16. PRICE CODE
Burst Test	Tim-Homogeneous Markov Chains	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE (include Area Code)	21c. OFFICE SYMBOL
J. K. Pugh	(619) 553-1632	Code 784

Biom. J. 35 (1993) 6, 677–688 Akademie Verlag

Sample Size Determination for Grouped Exponential Observations: A Cost Function Approach

KUNG-JONG LUI
DUANE STEFFEY
Department of Mathematical Sciences, College of Sciences,
San Diego State University

JAMIE K. PUGH
Naval Ocean Systems Center, San Diego

Summary

Calculating the required sample size for a desired power at a given type I error level, we often assume that we know the exact time of all subject responses whenever they occur during our study period. It is very common, however, in practice that we only monitor subjects periodically and, therefore, we know only whether responses occur or not during an interval. This paper includes a quantitative discussion of the effect resulting from data grouping or interval censoring on the required sample size when we have two treatment groups. Furthermore, with the goal of exploring the optimum in the number of subjects, the number of examinations per subject for test responses, and the total length of a study time period, this paper also provides a general guideline about how to determine these to minimize the total cost of a study for a desired power at a given α-level. A specified linear cost function that incorporates the costs of obtaining subjects, periodic examinations for test responses of subjects, and the total length of a study period, is assumed, primarily for illustrative purpose.

Key words: Interval censoring; Linear cost function; Maximum likelihood estimator; Sample size determination; Hazard rate.

1. Introduction

Calculating the sample size required to achieve a desired power at a fixed level of significance, we usually assume that we know the exact time of all subject responses that occur before the end of our study period (GROSS and CLARK, 1975; NARULA and LI, 1975; RASCH, 1977; EPSTEIN and SOBEL, 1953; GEORGE and DESU, 1974). It is very common, however, in practice that we only monitor subjects periodically and, therefore, we know only whether these subject responses have occurred during a given interval (KULLDORFF, 1961; CHENG and CHEN, 1988). We call this type of data grouped data, in which the information on the exact time of responses is unavailable. For example, when we want to study an asymptomatic chronic disease, continuous examination of all subjects in order to pinpoint the exact response time will often be practically difficult.

With regard to the optimal design of reliability tests, numerous researches have studied the properties of maximum likelihood estimators (MLEs) for grouped data under different model assumptions. KULLDORFF (1961) systematically laid down the fundamental theory for grouped data in exponential distributions. CHENG and CHEN (1988) discussed the conditions for the existence of the maximum likelihood estimator (MLE) and proposed an alternative estimator to the MLE under the Weibull model for grouped data. Furthermore, for a fixed number of examinations per subject, KULLDORFF (1961), NELSON (1977), WEI and BAU (1987), and WEI and SHAU (1987) also discussed for a variety of distributions the optimal interval length between any two consecutive examinations within the same subject in order to minimize the asymptotic variance of the MLE. None of these papers, however, addressed the optimal sample size determination when the number of subjects, the number of examinations per subject, and the length of a study period are permitted to vary simultaneously with the goal of minimizing the total costs incurred in conducting a study with a desired power $1-\beta$ at the α-level of significance.

In this paper, we first present the required number of subjects based on grouped exponential observations for a variety of parameter values in the situation where we have two treatment groups. This provides a quantitative assessment of the grouping effect on required sample sizes that are usually calculated with the assumption that the subject response times are known exactly. To generalize previous discussions on the optimality for grouped data (KULLDORFF, 1961; NELSON, 1977; WEI and BAU, 1987; WEI and SHAU, 1987), we consider a linear cost function that incorporates the costs of obtaining subjects, periodic examinations for test responses, and the per unit cost of maintaining the time period of a study. For a given power and a size requirement, we determine the required number of study subjects, the number of periodic examinations per subject, and the length of the study time period that minimize the total cost of a study.

2. Theory

Suppose that we have a completely randomized and balanced study that is designed to compare two treatment groups, of which each has n experimental subjects. Suppose also that the subject response times in the standard treatment and the experimental treatment groups are exponentially distributed with hazard rates λ_1 and λ_2, respectively. Because the standard treatment is replaced only when $\lambda_1 > \lambda_2$, we focus our discussion on testing the null hypothesis $H_0: \lambda_1 = \lambda_2$ versus the alternative hypothesis $H_a: \lambda_1 > \lambda_2$. Furthermore, we assume that subject withdrawal may occur in both treatment groups with withdrawal rates γ_1 and γ_2, respectively. To determine the response time for a given subject, we take K examinations that are equally spaced over the entire study period T_0.

Let Δ denote this common length in time between any two consecutive examinations; i.e., Δ simply equals T_0/K. Note that for a fixed study period T_0, if K were equal to ∞, Δ would equal 0. This corresponds to the situation in which we have no interval censoring in our data. Note also that the loss in the relative efficiency of using equidistant examination rather than the optimal grouping is generally not much. This is particularly true when the true values of λ_i, on which the optimal grouping strategy depends, are unknown and are thus required to be guessed, and when the number of examinations K is so large that $\lambda_i T_0/K < 0.8$ (KULLDORFF, 1961). Therefore, we will focus the following discussion on the equidistant case only.

Under the above assumptions, for a given subject j, where $j = 1, 2, \ldots, n$, in treatment group i ($i = 1, 2$), the probability that the response time R_{ij} falls in the interval $((k_{ij}-1)\Delta, k_{ij}\Delta)$, where $k_{ij} = 1, 2, \ldots, K$, equals

$$P(R_{ij} \in ((k_{ij}-1)\Delta, k_{ij}\Delta)) = \exp(-\lambda_i(k_{ij}-1)\Delta)(1-\exp(-\lambda_i\Delta)). \tag{1}$$

Similarly, for the subject j, the probability that withdrawal time W_{ij} falls in the interval $((k_{ij}-1)\Delta, k_{ij}\Delta)$ equals

$$P(W_{ij} \in ((k_{ij}-1)\Delta, k_{ij}\Delta)) = \exp(-\gamma_i(k_{ij}-1)\Delta)(1-\exp(-\gamma_i\Delta)). \tag{2}$$

Therefore, the general likelihood L_i for n subjects in group i ($i = 1, 2$) can be written as

$$\begin{aligned} L_i = \prod_{j=1}^{n} & [\exp(-\lambda_i(k_{ij}-1)\Delta)(1-\exp(-\lambda_i\Delta))\exp(-\gamma_i(k_{ij}-1)\Delta)]^{\delta_{ij}} \\ & [\exp(-\gamma_i(k_{ij}-1)\Delta)(1-\exp(-\gamma_i\Delta))\exp(-\lambda_i(k_{ij}-1)\Delta)]^{\delta^*_{ij}} \\ & [\exp(-(\lambda_i+\gamma_i)K\Delta)]^{1-\delta_{ij}-\delta^*_{ij}} \end{aligned} \tag{3}$$

where

$$\delta_{ij} = \begin{cases} 1 \text{ if subject } j \text{ in group } i \text{ responded before withdrawal and the end of our study period } T_0, \\ 0, \text{ otherwise,} \end{cases}$$

and

$$\delta^*_{ij} = \begin{cases} 1 \text{ if subject } j \text{ in group } i \text{ withdrew before response and the end of our study period } T_0, \\ 0, \text{ otherwise.} \end{cases}$$

On the basis the above likelihood (3), we can obtain the MLE of λ_i by solving the following equation:

$$\frac{\partial \log L_i}{\partial \lambda_i} = \sum_j -[\delta_{ij} + \delta^*_{ij}]\,[k_{ij}-1]\Delta + \delta_{ij}\left[\frac{\Delta \exp(-\lambda_i \Delta)}{(1-\exp(-\lambda_i \Delta))}\right] - \\ - K\Delta(1-\delta_{ij}-\delta^*_{ij}) = 0. \tag{4}$$

This leads the MLE of λ_i to be

$$\hat{\lambda}_i = \frac{1}{\Delta}\log\left(1 + \frac{\sum_j \delta_{ij}}{\sum_j [(\delta_{ij}+\delta^*_{ij})(k_{ij}-1) + K(1-\delta_{ij}-\delta^*_{ij})]}\right). \tag{5}$$

Furthermore, because $\dfrac{-\partial^2 \log L}{\partial \lambda_i^2} = \sum_j \delta_j \left[\dfrac{\Delta^2 \exp(-\lambda_i \Delta)}{(1-\exp(-\lambda_i \Delta))^2}\right]$, and because

$$E(\delta_j) = \frac{\lambda_i}{\lambda_i + \gamma_i}\{1 - \exp[-(\lambda_i + \gamma_i)T_0]\},$$

the asymptotic variance of $\log(\hat{\lambda}_i)$ is

$$\mathrm{Var}(\log(\hat{\lambda}_i)) = \frac{V(\lambda_i, \gamma_i, T_0, K)}{n}, \tag{6}$$

where

$$V(\lambda_i, \gamma_i, T_0, K) = \left\{\frac{\lambda_i}{\lambda_i+\gamma_i}[1-\exp(-(\lambda_i+\gamma_i)T_0)]\right\}^{-1} \\ \cdot \frac{(1-\exp(-\lambda_i T_0/K))^2}{\exp(-\lambda_i T_0/K)(\lambda_i T_0/K)^2}.$$

Note that when $\gamma_i = 0$ (i.e., there is no subject withdrawal), the MLE $\hat{\lambda}_i$ and the asymptotic variance in formulae (5) and (6) reduce to those derived by KULLDORFF (1961) and NELSON (1977). Note also that for a fixed ratio $\lambda_i T_0$, between the study period and the mean response times, variance formula (6) is an increasing function of the loss rate γ_i and is a decreasing function of the number K of examinations per subject. Furthermore, if K increased to ∞,

$$\frac{(1-\exp(-\lambda_i T_0/K))^2}{\exp(-\lambda_i T_0/K)(\lambda_i T_0/K)^2}$$

would decrease to 1. Therefore, the variance formula (6) will reduce to the same as that when we know the exact response times that occur before T_0.

Consider testing the null hypothesis $H_0: \lambda_1 = \lambda_2$ versus the alternative hypothesis $H_a: \lambda_1/\lambda_2 = R > 1$ with a desired power $1-\beta$ at α-level. In developing sample size formulae, as frequently assumed elsewhere (TAULBEE and SYMONS, 1983; GROSS, HUNG, CANTOR, and CLARK, 1987), we rely on the asymptotic normality of the MLE. On the the basis of the above formula (6), for a fixed study period T_0 and for a given number K of examinations per subject, the required number of subjects, n, is then given by the smallest integer larger than

$$\frac{(Z_\alpha\sqrt{V(\bar{\lambda}, \gamma_1, T_0, K) + V(\bar{\lambda}, \gamma_2, T_0, K)} + Z_\beta\sqrt{V(\lambda_1, \gamma_1, T_0, K) + V(\lambda_1/R, \gamma_2, T_0, K)})^2}{(\log(R))^2} \tag{7}$$

where $\bar{\lambda} = \lambda_1\,(1+1/R)/2$, and where Z_α and Z_β are the upper 100αth and 100βth percentiles of a standard normal distribution, respectively.

Let C_1 and C_2 denote the costs per subject and per examination, respectively. Furthermore, let C_3 denote the cost per unit time during our study period T_0. Therefore, the total cost for a treatment group in the study under consideration here is given by $2C_1 \times n + 2C_2 \times n \times K + C_3 \times T_0$. To detect the treatment effect $R = \lambda_1/\lambda_2 > 1$ at α-level with a desired power $1-\beta$, we want to find n, K and T_0 that minimize the above linear cost function subject to the following constraints: that (i) n and K are integers, and (ii)

$$\frac{Z_\alpha\sqrt{V(\bar{\lambda}, \gamma_1, T_0, K) + V(\bar{\lambda}, \gamma_2, T_0, K)} - \log(R)\sqrt{n}}{\sqrt{V(\lambda_1, \gamma_1, T_0, K) + V(\lambda_1/R, \gamma_2, T_0, K)}} = -Z_\beta. \tag{8}$$

We apply the IMSL subroutine DNCONF to obtain these optimal solutions as follows: We employ a sequential procedure in which we first solve the optimization problem for minimizing the above linear cost function subject to the constraint (8), but without imposing the integer constraints. Then, the value of K is fixed, in turn, at the greatest integer smaller than unconstrained value, and the smallest integer greater than the unconstrained value. The subsequent two-dimensional problem (involving n and T_0) is solved, again without constraining n to be integer-valued. The value for n is then fixed at the smallest integer larger than its unconstrained value, and the value of T_0 is found by solving the subsequent one-dimensional problem. This procedure is repeated with the roles of K and n reversed. Of the resulting four sets of values for (K, n, T_0), the set with the smallest cost is selected as our final optimal solution.

3. Results

To study the possible loss of efficiency resulting from grouping, we summarize the required sample sizes for the power of 0.90 at 0.05 level calculated by using sample size formula (7). Tables 1 and 2 present the results for the hazard rate λ_1 in the standard treatment group ranging from 0.015 to 0.030, the fixed time

study period T_0 ranging from 50 to 400, the treatment effect R, ranging from 2 to 4, the number of observations K per subject ranging from 1 to infinity, and for the common loss rate γ equal to 0.0 and 0.01. For example, when the loss rate $\gamma = 0.00$, the hazard rate $\lambda_1 = 0.015$, the study period $T_0 = 50$, and R from 2 to 4 (Table 1), taking one or two observations per subject is almost as efficient as taking infinitely many observations per subject (i.e., knowing the exact time of test response), as evidenced by the fact that the required sample size is virtually constant over all values of K. This is true for the same configuration as above with the loss rate equal to 0.01 (Table 2). By contrast, when the hazard rate is $\lambda_1 = 0.030$ and the study period increases to 400, the loss of efficiency due to grouping is substantial and can remain as large as 20 to 30% (Tables 1 and 2), even when taking as many as 5 examinations per subject.

Table 1

Required sample sizes for the power of 0.90 at 0.05 level (one-sided test) based on grouped exponential observations with hazard rate in the standard treatment group, λ_1, ranging from 0.015 to 0.030, the fixed time study period T_0, ranging from 50 to 400, the treatment effect R $(=\lambda_1/\lambda_2)$, ranging from 2 to 4, and the number of examinations per subject, K, ranging from 1 to ∞ at the loss rate, γ, equal to 0.00 per unit of time.

λ_1	T_0	R	$K=1$	2	3	4	5	10	∞
0.015	50	2	89	87	87	87	87	87	87
		3	41	41	41	41	41	41	41
		4	29	29	29	29	29	29	29
	100	2	61	56	56	55	55	55	55
		3	27	25	25	25	25	25	25
		4	19	18	18	17	17	17	17
	200	2	63	46	43	42	42	41	41
		3	25	19	19	18	18	18	18
		4	16	13	13	12	12	12	12
	400	2	190	57	45	41	39	37	37
		3	65	22	18	17	16	16	15
		4	39	14	12	11	11	10	10
0.025	50	2	66	62	62	61	61	61	61
		3	30	28	28	28	28	28	28
		4	21	20	20	20	20	20	20
	100	2	58	47	45	44	44	44	44
		3	24	20	20	19	19	19	19
		4	16	14	13	13	13	13	13
	200	2	119	50	43	40	39	38	37
		3	43	20	18	17	16	16	16
		4	26	13	11	11	11	11	10
	400	2	2194	116	62	49	44	38	36
		3	693	41	23	19	17	15	15
		4	396	25	14	12	11	10	10

(Continuation) Table 1

λ_1	T_0	R	$K=1$	2	3	4	5	10	∞
0.030	50	2	61	56	56	55	55	55	55
		3	27	25	25	25	25	25	25
		4	19	18	18	17	17	17	17
	100	2	63	46	43	42	42	41	41
		3	25	19	19	18	18	18	18
		4	16	13	13	12	12	12	12
	200	2	190	57	45	41	39	37	37
		3	65	22	18	17	16	16	16
		4	39	14	12	11	11	10	10
	400	2	9227	188	77	56	48	39	36
		3	2888	64	28	21	19	16	15
		4	1651	38	17	13	12	10	10

Table 2

Required sample sizes for the power of 0.90 and 0.05 level (one-sided test) Based on grouped exponential observations with hazard rate, λ_1, in the standard treatment group ranging from 0.015 to 0.030, the fixed time study period T_0, ranging from 50 to 400, the treatment effect R $(=\lambda_1/\lambda_2)$, ranging from 2 to 4, and the number of examinations per subject, K, ranging from 1 to ∞ at the loss rate, γ, equal to 0.01 per unit of time.

λ_1	T_0	R	$K=1$	2	3	4	5	10	∞
0.015	50	2	111	108	108	108	108	108	108
		3	51	51	51	51	51	51	51
		4	36	36	36	36	36	36	36
	100	2	88	82	81	80	80	80	80
		3	39	37	37	37	37	37	37
		4	27	26	26	26	26	26	26
	200	2	107	79	74	73	72	71	71
		3	44	34	33	32	32	32	32
		4	29	23	23	22	22	22	22
	400	2	344	105	84	77	74	71	70
		3	121	43	36	34	33	31	31
		4	73	28	24	23	22	21	21
0.025	50	2	80	76	75	75	75	75	75
		3	36	35	35	35	34	34	34
		4	25	24	24	24	24	24	24
	100	2	80	65	62	61	61	60	60
		3	34	28	27	27	27	27	27
		4	22	19	19	19	19	18	18
	200	2	176	76	64	61	59	57	56
		3	64	31	27	26	26	25	25
		4	40	20	18	18	17	17	17
	400	2	3201	176	94	75	68	59	56
		3	1019	64	37	31	28	25	24
		4	584	39	24	20	19	17	16

(Continuation) Table 2

λ_1	T_0	R	$K=1$	2	3	4	5	10	∞
0.030	50	2	74	69	68	67	67	67	67
		3	33	31	31	31	31	30	30
		4	23	22	21	21	21	21	21
	100	2	84	62	58	57	56	56	55
		3	34	27	25	25	25	25	24
		4	22	18	17	17	17	17	17
	200	2	267	81	64	59	57	54	53
		3	93	32	27	25	24	23	23
		4	56	21	18	17	16	16	15
	400	2	12675	266	111	81	69	57	53
		3	3981	93	42	32	28	24	23
		4	2279	56	26	21	19	16	15

Table 3

Optimal results about length of study period, T_0, sample sizes of subjects, n, and the number of examinations per subject, K, for the power of 0.90 at 0.05 Level (one-sided test) based on grouped exponential observations with hazard rate, λ_1, in the standard treatment group ranging from 0.015 to 0.030, and the treatment effect R $(=\lambda_1/\lambda_2)$, ranging from 2 to 4 at the loss rate, γ, equal to 0.00.

		C_3/C_2	0			2			10		
C_1/C_2	λ_1	R	T	n	K	T	n	K	T	n	K
1.0	0.015	2	129.7	57	1	72.5	70	1	35.7	114	1
		3	135.6	24	1	53.2	39	1	25.2	72	1
		4	136.0	16	1	47.3	30	1	21.2	60	1
	0.025	2	77.8	57	1	52.1	64	1	27.2	95	1
		3	81.3	24	1	40.6	33	1	19.4	58	1
		4	81.6	16	1	36.0	25	1	16.3	48	1
	0.030	2	64.9	57	1	46.8	62	1	24.8	89	1
		3	67.8	24	1	37.3	31	1	17.6	54	1
		4	68.0	16	1	31.7	24	1	15.0	44	1
5.0	0.015	2	174.7	46	2	102.5	60	1	59.2	79	1
		3	197.0	19	2	83.7	29	1	42.9	46	1
		4	176.5	13	2	77.3	21	1	36.7	37	1
	0.025	2	104.8	46	2	69.7	58	1	43.5	70	1
		3	118.2	19	2	62.8	26	1	31.9	39	1
		4	105.9	13	2	54.7	19	1	28.4	30	1
	0.030	2	87.4	46	2	58.1	58	1	39.4	67	1
		3	98.5	19	2	52.3	26	1	29.8	36	1
		4	88.3	13	2	50.5	18	1	25.8	28	1

(Continuation) Table 3

		C_3/C_2	0			2			10		
C_1/C_2	λ_1	R	T	n	K	T	n	K	T	n	K
10.0	0.015	2	220.2	42	3	145.9	48	2	76.5	68	1
		3	285.4	17	3	104.7	26	1	57.2	37	1
		4	302.4	11	3	91.2	19	1	49.4	29	1
	0.025	2	132.1	42	3	94.6	47	2	54.0	63	1
		3	171.2	17	3	85.0	21	2	42.6	32	1
		4	181.4	11	3	68.5	17	1	36.0	25	1
	0.030	2	110.1	42	3	87.4	46	2	46.8	62	1
		3	142.7	17	3	80.9	20	2	37.3	31	1
		4	151.2	11	3	73.1	14	2	33.7	23	1
20.0	0.015	2	262.3	40	4	191.8	43	3	107.5	54	2
		3	285.4	17	3	141.7	21	2	74.7	31	1
		4	302.4	11	3	146.2	14	2	63.5	24	1
	0.025	2	157.4	40	4	132.1	42	3	77.5	50	2
		3	171.2	17	3	121.6	18	3	59.8	25	2
		4	181.4	11	3	87.7	14	2	50.1	20	1
	0.030	2	131.1	40	4	110.1	42	3	68.4	49	2
		3	142.7	17	3	101.3	18	3	53.6	24	2
		4	151.2	11	3	88.3	13	2	46.4	18	2

Table 4

Optimal results about length of study period, T_0, sample sizes of subjects, n, and the number of examinations per subject, K, for the power of 0.90 at 0.05 level (one-sided test) based on grouped exponential observations with hazard rate, λ_1, in the standard treatment group ranging from 0.015 to 0.030, and the treatment effect R $(=\lambda_1/\lambda_2)$, ranging from 2 to 4 at the mean loss rate, γ, equal to 0.01 per unit of time.

		C_3/C_2	0			2			10		
C_1/C_2	λ_1	R	T	n	K	T	n	K	T	n	K
1.0	0.015	2	97.9	88	1	68.1	96	1	34.9	136	1
		3	99.7	39	1	51.7	50	1	24.8	82	1
		4	123.0	26	1	45.5	38	1	21.4	66	1
	0.025	2	61.5	76	1	49.7	80	1	26.9	108	1
		3	65.8	33	1	39.3	40	1	19.2	64	1
		4	73.1	22	1	34.7	30	1	16.3	52	1
	0.030	2	59.3	72	1	43.9	76	1	24.2	101	1
		3	62.6	31	1	36.1	37	1	17.5	59	1
		4	61.6	21	1	31.1	28	1	15.1	47	1
5.0	0.015	2	97.9	88	1	90.3	89	1	56.1	104	1
		3	99.7	39	1	75.8	42	1	41.5	57	1
		4	123.0	26	1	71.0	30	1	36.4	44	1
	0.025	2	102.1	64	2	61.5	76	1	42.2	85	1
		3	95.4	28	2	58.2	34	1	31.9	45	1
		4	73.1	22	1	53.9	24	1	27.4	35	1
	0.030	2	82.4	61	2	52.1	73	1	37.9	80	1
		3	62.6	31	1	47.4	33	1	28.6	42	1
		4	61.6	21	1	45.6	23	1	25.2	32	1

(Continuation) Table 4

		C_3/C_2	0			2			10		
C_1/C_2	λ_1	R	T	n	K	T	n	K	T	n	K
10.0	0.015	2	141.6	77	2	123.9	78	2	70.2	95	1
		3	147.6	34	2	88.7	40	1	53.7	49	1
		4	166.9	23	2	85.1	28	1	47.5	37	1
	0.025	2	102.1	64	2	84.7	65	2	49.7	80	1
		3	95.4	28	2	79.6	29	2	41.3	39	1
		4	95.5	19	2	60.8	23	1	36.7	29	1
	0.030	2	82.4	61	2	71.3	62	2	43.9	76	1
		3	73.7	27	2	63.8	28	2	36.1	37	1
		4	80.1	18	2	65.4	19	2	33.0	27	1
20.0	0.015	2	141.6	77	2	123.9	78	2	92.3	83	2
		3	147.6	34	2	121.3	35	2	67.3	44	1
		4	166.9	23	2	124.1	24	2	61.8	32	1
	0.025	2	108.4	61	3	108.4	61	3	72.1	67	2
		3	95.4	28	2	79.6	29	2	54.6	33	2
		4	110.5	18	3	77.4	20	2	45.0	26	1
	0.030	2	89.7	58	3	89.7	58	3	61.1	64	2
		3	95.8	25	3	73.7	27	2	52.3	30	2
		4	129.4	16	4	80.1	18	2	41.4	24	1

We next address the problem of minimizing total study cost subject to integer constraints on n and K and the inference constraint (8) for a given power of 0.90 at 0.05 level of significance. Tables 3 and 4 summarize the results by giving the study periods, T_0, the required number of subjects, n, and the number of examinations per subject, K, for the subject/examination unit cost ratio, C_1/C_2, ranging from 1 to 20, the time/examination unit cost ratio, C_3/C_2, ranging from 0 to 10, the basic hazard rate, λ_1, ranging from 0.015 to 0.030, the treatment effect, R, ranging from 2 to 4, and loss rates, γ, equal to 0.00 and 0.01. For example, when $C_1/C_2=10$, $C_3/C_2=2.0$, $\lambda_1=0.030$, and $R=3$, the optimal study period T_0, the required number of subjects, n, and the number of examinations K per subject are 80.9, 20, and 2 (Table 3). In this situation but for the loss rate $\gamma=0.01$ (Table 4), these optimal solutions for T_0, n, and K become 63.8, 28, and 2, respectively.

4. Discussion

If we know the exact times of the test responses (or equivalently, the number of examinations, K, per subject is infinite), then since

$$\mathrm{Var}(\log(\hat{\lambda}_i))=\left\{\frac{n\lambda_i}{\lambda_i+\gamma_i}\left[1-\exp(-(\lambda_i+\gamma_i)T_0)\right]\right\}^{-1}$$

is a decreasing function of the study time period T_0, the required sample size always decreases as T_0 increases (Tables 1 and 2). This is not true, however, for grouped data. In fact, for a fixed number of observations k per subject, increasing the length of study period T_0 alone can conversely lead to requiring a larger sample size of subjects (Tables 1 and 2). This increase can be substantial, especially when the ratio of the length of study period to the mean response time $\lambda_i T_0$ is large and K is small. Therefore, when we have grouped data and when the underlying test responses have a short mean lifetime, we should consider increasing the number of observations, K, per subject before increasing the study period to improve the precision of the MLE. On the other hand, if $\lambda_i T_0$ or T_0 is small (Tables 1 and 2), the required number of subjects for $K=2$ is essentially equal to that for $K=\infty$.

From both Tables 3 and 4, the optimal number of examinations, K, increases with an increase of the relative costs C_1/C_2. This is consistent with our intuition, because if the cost of obtaining a subject were relatively larger than that of obtaining an examination, it would be certainly wise to increase the number of examinations per subject in order to reduce the required number of subjects to reach a desired power. The optimal number of examinations, K, however, decreases with an increase of C_3/C_2. This can be interpreted as a result that when the cost of per unit study period is relatively high, it may be more economical to increase the number of subjects and reduce the study period T_0. Therefore, in this situation in which T_0 is small, as noted before, taking one or two examinations is as efficient as taking infinitely many ($K=\infty$) examinations. In other words, increasing the number of examinations per subject, K, does not significantly reduce the asymptotic variance. Note also that even when the relative cost C_1/C_2 is as large as 20, the optimal number of examinations per subject, K, is not bigger than 4 (Tables 3 and 4). This suggests that, in practice, we may seldom need to take more than 4 observations per subject.

When the loss rate γ is greater than 0, there is nonzero probability that a given subject will drop out before the end of our study and this probability will increase with the length of the study period. Therefore, comparing the results in Table 3 ($\gamma=0$) with those in Table 4 ($\gamma=0.01$), we will increase the required number of subjects rather than increase the length of the study period in order to compensate for the missing information resulting from those drop-out subjects. In fact, we can see that the optimal length of our study periods, T_0 in Table 4, generally are shorter than the corresponding ones in Table 3 and so are the optimal number of examinations per subject.

In summary, in this paper, we have derived a general sample size formula for the required number of subjects for grouped exponential data. We have quantitatively studied the grouping effect directly on the required sample sizes of subjects in a variety of situations. We also have included a discussion on the optimal sample allocation and the optimal length of the study period under a linear cost function.

Acknowledgement

The authors wish to thank Dr. Thomas Bonifield of the U.S. Public Health Service for his encouragement. This work was supported by Navy Computer Telecommunications Station, Funding Document No. N6316591WR60001.

References

CHENG, K. F. and CHEN, C. H., 1988: Estimation of the Weibull parameters with grouped data. Comm. Statist. Theor. Meth. **17**, 325–341.

EPSTEIN, B. and SOBEL, M., 1953: Life testing. J. Amer. Statist. Assoc. **48**, 486–502.

GEORGE, S. L. and DESU, M. M., 1974: Planning the size and duration of a clinical trial studying the time to some critical event. J. Chronic Diseases **27**, 15–24.

GROSS, A. J. and CLARK, V. A., 1975: *Survival Distributions: Reliability Applications in the Biomedical Sciences.* Wiley, New York.

GROSS, A. J., HUNG, H. H., CANTOR, A. B., and CLARK, B. C., 1987: Sample size determination in clinical trials with an emphasis on exponentially distributed responses. Biometrics **43**, 875–883.

KULLDORFF, G., 1961: *Estimation from Grouped and Partially Grouped Samples.* Wiley, New York.

NARULA, S. C. and LI, F. S., 1975: Sample size calculations in e..ponential life testing. Technometrics **17**, 229–231.

NELSON, W., 1977: Optimum demonstration tests with grouped inspection data from an exponential distribution. IEEE Transactions on Reliability **26**, 226–231.

RASCH, D., 1977: Sample size determination of estimating the parameter of an exponential distribution. Biom. J. **19**, 521–528.

TAULBEE, J. D. and SYMONS, M. J., 1983: Sample size and duration for cohort studies of survival time with covariables. Biometrics **39**, 351–360.

WEI, D. and BAU, J. J., 1987: Some optimal designs for grouped data in reliability demonstration tests. IEEE Transactions on Reliability **36**, 600–604.

WEI, D. and SHAU, C. K., 1987: Fitting and optimal grouping on Gamma reliability data. IEEE Transactions on Reliability **36**, 595–599.

Received June 1992

Dr. K.-J. LUI
Department of Mathematical Sciences
College of Sciences
San Diego State University
San Diego, CA 92182-0314
U.S.A.